UNIVERSITY OF CALIFORNIA, BERKELEY

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

Narrative for Tenure

Prof. Teresa Head-Gordon

Department of Bioengineering

University of California, Berkeley

May 14, 2003

Introduction

I am an assistant professor in the Department of Bioengineering (BE) in the College of Engineering at UC Berkeley, working in the area of computational biology as one of BE's four major thrust areas (http://www-bioeng.berkeley.edu/research.html). I am also a faculty staff scientist in the QB3 Institute (the new California Institute for Quantitative Biomedical Research), and department head of the Computational Structure group in the Physical Biosciences Division at Lawrence Berkeley National Laboratory. I am a faculty member of three graduate programs: Joint Graduate Group in Bioengineering, Biophysics graduate group, and Applied Sciences and Technology (AS&T) graduate group, and a steering member of the Computational Sciences and Engineering undergraduate program at UC Berkeley.

My research program encompasses the development of general computational methodologies and modeling applied to biology in the areas of water and aqueous hydration, protein folding and structure prediction, and there is a growing experimental component to my laboratory in these same areas. I have over 50 publications and several research grants reflecting my activity and accomplishments in these areas. I have also been involved in local and national service, education, and training, which extends to promoting and developing the blueprint for computational biology and biophysical research for the future. In the following pages I provide a detailed narrative on research, teaching, and service during the tenure track period between 1/2001-5/2003.

Research Narrative

Theory/Experiment: Hydration Forces in Protein Folding.

The experimental repertoire of protein structure determination is highly developed, and is typical of research conducted at the various protein crystallography beamlines throughout the world. Even though it is well appreciated that water environment is a vital determinant of protein tertiary structure and stability, the experimental tools available to characterize aqueous environment in terms of structure and forces are comparatively minimal at present. We have combined wide-angle x-ray and neutron solution diffraction experiments, simulations, and theory, to determine hydration structure and forces in both early and late stages of protein folding through a well defined model system of amino acid monomers as a function of concentration. This work has appeared in five publications (7, 11, 20, 22, 23), including a feature article in J. Phys. Chem. (22).

We believe that our experiments support the view that small lengthscale hydration physics (Pratt Chandler theory) is operative for the folding of globular proteins based on model studies of amino acid monomers in solution. We have used these potentials of mean force (pmf) in our protein structure prediction work (see below) as a description of hydrophobic solvation, and have shown how the pmf contributed to the successful prediction of one of the most difficult targets in CASP4. We are also further characterizing the amino acid monomer solutions by quasi-elastic neutron scattering to determine timescales of water motion near exposed amino acid side chains (dilute solution), and that for pure hydration water (high concentration), of both hydrophobic and hydrophilic residues (manuscript in preparation).

The experimentally determined solvation of amino acid monomers are being extended to real protein chains by considering the role of hydration in stabilizing molten globule intermediates of various lysozymes and αlactalbumin. Currently we are setting up neutron and x-ray solution scattering experiments on labeled versions (deuterium and selenium) of both the native and molten globule forms of the lysozyme fold class, as well as mutant sequences defined as all leucine or all methionine hydrophobic cores (synthesis in collaboration with David King at UCB). These experiments will determine differences in intensity that arise from different structural organization of the hydrophobic core under native and molten globule conditions. Our hypothesis is that molten globule states have a well-defined tertiary structure when water is considered as the “twenty-first” amino acid, with well-defined water positions that are incorporated between hydrophobic side chains.

Pure Water Structure:

We reported a new, high-quality x-ray scattering experiment on pure ambient water using synchrotron beam line 7.3.3 at the Advanced Light Source at LBNL and showed using more advanced theoretical analysis that the older x-ray curves support a family of $g_{OO}(r)$’s that are different than that from the ALS experiment. Two J. Chem. Phys. papers (one experiment, one theory) that describe this highly accurate water structure determination was published in 2000 (17,18), and the data has made its way into research groups world-wide. We also provided a review of water structure for Chemical Reviews in 2002 (11).

In our most recent work, we obtained high-quality x-ray scattering experiments on pure water taken over a temperature range of 2°C to 77°C. The ALS x-ray scattering intensities are *qualitatively* different in trend of maximum intensity over this temperature range compared to older x-ray experiments. While the common procedure is to report both the intensity curve and radial distribution function(s), the proper extraction of the real-space pair correlation functions from the experimental scattering is very difficult due to uncertainty introduced in the experimental corrections, the proper weighting of OO, OH, and HH contributions, and numerical problems of Fourier transforming truncated data in Q-space. Instead, in collaboration with M. Krack and M. Parrinello, we considered the direct calculation of x-ray scattering spectra using electron densities derived from density functional theory based on real-space configurations generated with classical water models. We found that the TIP4P, TIP5P and polarizable TIP4P-Pol2 water models, with DFT-LDA densities, show very good agreement with the experimental intensities, and TIP4P-Pol2 in particular shows quantitative agreement over the full temperature range. The resulting radial distribution functions from TIP4P-Pol2 provide the current best benchmarks for real-space water structure over the biologically relevant temperature range studied here. This work was published in PCCP in 2003 (5).

We hope to continue our high standard in determining water structure by considering other areas of waters phase diagram (supercritical and supercooled for example), as well as higher accuracy in extending these experiments to very high Q. We will also expand our theoretical work in developing models and simulation methods for calculation of the electron density in the condensed phase for water.

Global Optimization Approaches to Protein Structure Prediction:

The protein folding problem in its most pragmatic guise is to predict the full three-dimensional structure of the protein molecule given a protein-solvent potential or free energy surface, and the amino acid sequence as input. The “rugged landscape” topography of this surface defines the underlying difficulty the native structure minimum, usually the global minimum, must be discriminated from other minima whose number rises exponentially with the number of amino acids in the sequence. Furthermore, this energy surface is difficult to model reliably in a global sense, i.e. to ensure that all misfolds are higher in energy than the correctly folded conformation.

Given the difficulty of these two obstacles, a pragmatic alternative viewpoint is to diminish the emphasis on proteins as strictly physical systems, but instead to exploit analogies in databases of existing protein structures. The most successful methods at present are those that can most effectively use information from the sequence and structure of known proteins to form some type of structural template for predicting tertiary structure of unknown targets. For targets, or portions of targets where this information is unavailable, these methods may be somewhat less successful than those that rely more on generically applicable physical principles, and optimization approaches may be particularly important.

Our group (in collaboration with Byrd and Schnabel, U. Colorado, Boulder) has developed a new fold method called Stochastic Perturbation with Soft Constraints (SPSC), which uses information from known proteins to predict secondary structure, but not in the tertiary structure predictions or in generating the terms of the physics-based energy function. We have developed a number of global optimization approaches, including SPSC, potential smoothing, the antlion strategy, and most recently C-walking. We are also exploring use of “implicit” hydration potentials between amino acids in solution derived from experiment/simulation (described above) as new energy functions for structure prediction. Much if the last two years has been devoted to development of the necessary infrastructure to obtain blind prediction results with our method and energy function in the 4th and 5th Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition. This work has resulted in five publications (2, 10, 14, 15, 24), including the description of our CASP4 results in 2002 (10).

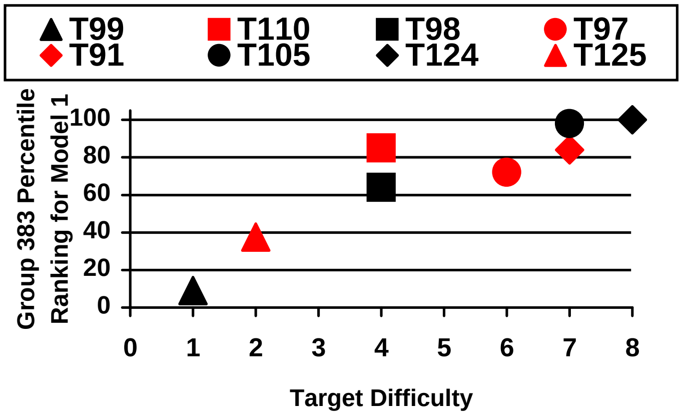

Figure 1. Difficulty of CASP4 targets as rated by the CASP4 organizers vs. the percentile ranking of our groups (Group 383) submissions using our SPSC global optimization method. The percentile ranking of our models (based on GDT_TS scores) generally increases with target difficulty. Percentile raking with respect to target difficulty for all models submitted.

In Figure 1, we plot the comparative difficulty of each protein we attempted in CASP4 (Group 383) versus the comparative accuracy of our prediction to that of other groups. Figure 1 shows that as the difficulty of the targets increases, the percentile of our models generally increases as well. Target 124 (Phospholipase C beta C-terminus, turkey) was considered to be one of the most difficult targets of the conference by the CASP4 organizers. It was also a difficult target from an optimization point of view, with 242 amino acids, 4102 atoms, and over 12,000 cartesian coordinates. Our submission had the best GDT_TS score of all submissions for this target, and on an absolute scale was a highly accurate fold prediction, with an overall RMSD of 8.46Å. The results for this target show the value of a computational optimization method that does not rely on known protein structures for predicting proteins with new folds. In CASP5 we submitted *ab initio* predictions on 20 complete targets, some as large as 420 amino acids, with complicated fold topologies. Our ranking in the new fold/fold recognition category was 13th-15th (depending on metric used) out of 154 groups evaluated (see Group 271 at http://www.russell.embl.de/casp5/).

Minimalist Models for Protein Folding and Design

While the experimental effort in structural genomics is partly focused on providing new fold classifications, computation and theory should play a complementary role of completing structural, kinetic, and thermodynamic information across whole genomes. In this case, reduction in computational complexity of the model will be necessary but retaining physical-chemical connections to experiment will be vital. Analytic theories of protein folding have provided the criteria for successful folding models, while our recent work seeks to design protein models consistent with these criteria while establishing stronger quantitative connections to experiments.

The philosophy of these models is to avoid the use of Gō potentials, but instead to make tertiary folding predictive using physically motivated potentials, and therefore reestablish the connection between free energy landscapes and amino acid sequence (the original protein folding problem). We have now completed multiple studies of a minimalist protein folding model addressing issues of protein sequence design, the role of solvation and interaction complexity in protein folding models, the ability to design and validate the folding of complex topologies, longer protein chains, and most recently in regards to protein engineering (phi-value analysis) studies. In combination we believe they indicate the utility of minimalist modeling as a feasible approach to providing complete structural, thermodynamic, and kinetic information on any protein of interest.

This work has resulted in eight publications (3,4,6,8,16,19,26,29), including a *Curr. Opin. Struct. Biol.* summarizing minimalist model approaches in protein folding and design in 2003 (4). Currently we are adapting the model to include an orientational hydrogen-bonding potential to reduce our dependence on secondary structure knowledge. We are extending the application to studies of protein aggregation (see below). Ultimately these models will be interfaced with complex all-atom potentials for eventual use in multiscale modeling.

Models of the Condensed Phase in Ab Initio Molecular Orbital Theory

Reaction field (RF) methods are one of the simplest approaches to modeling the condensed phase since only the solute is treated in microscopic detail. We have developed a new quantum reaction field model that does not require empirical specification of the cavity shape and size. When the solute wavefunction is optimized under the boundary condition that it is fully contained within the cavity, reaction field stabilization and repulsive kinetic energy compression are the competing forces that allow an optimal cavity size to be determined. A recent new model relaxes the strict separation of solute and solvent electron densities to allow leakage of the solute wavefunction into the surrounding dielectric with a density functional penalty that is a function of solvent and solute electron density. This work has resulted in three publications (25, 27, 28).

Most recently, my group is exploring the use of this model as a way to manifest the effects of *pressure* on electronic structure stability and other properties of potential energy surfaces. We also are developing new *ab initio* approaches to condensed phase simulation using Monte Carlo (MC) and eventually molecular dynamics (MD). In collaboration with the IBM Blue Gene team, we propose the development of a hybrid energy MC approach that uses a cheap energy function to provide configurations for the Metropolis sampling of an expensive *ab initio* energy. This has resulted in one paper in *J. Comp. Chem.* in 2003 (1), in which we describe a new ergodic sampling method known as Cool-walking, but which may be helpful in convergence issues for the hybrid energy MC approach.

Modeling Chemical Bonding Effects for Protein Electron Crystallography.

Electron crystallography is a relatively new approach for structure determination for proteins, with increasing ability to generate atomic resolution structures of biologically significant proteins. During the refinement of electron crystallography data, a superposition of free atom form factors is used to represent the electrostatic potential. However, scattering of electrons is affected by the distribution of valence electrons that participate in chemical bonding and thus change the electrostatic shielding of the nucleus. This effect is particularly significant for low-angle scattering, and therefore can be substantial in the study of proteins by electron crystallography. We have investigated the magnitude of chemical bonding effects for a representative collection of protein fragments and a model ligand for nucleotide binding proteins within the resolution range generally used in determining protein structures by electron crystallography. Electrostatic potentials were calculated by *ab initio* methods for both the test molecules and for superpositions of their free atoms. Differences in scattering amplitudes can be well over 10% in the resolution range below 5 Å and are especially large in the case of ionized side chains and ligands. It is quite clear that better use of the low-resolution data for refinement can be realized by accounting for chemical bonding effects.

Our approach is to replace the free atom form factors with transferable form factors that incorporate chemical bonding effects. These ideas have been advanced by us recently by consideration of different atomic and molecular fragments to reproduce the molecular electrostatic potential of different conformations of N-acetyl alanine-methylamide (NAAMA) with an acceptable degree of error as measured by conventional R-factors used in the refinement procedure common in crystallography. We have examined transferable fragments that incorporate increasingly more complete descriptions of molecular bonding with diminishing accuracy in geometric fit to the target molecule: single atoms in molecules, bonded atoms in molecules, and selected larger functional groups. In the resolution range between 2.5-25.0Å, we find that the fairly straightforward use of single atoms in molecules reduces the calculated R-factors by 5-15% over a free-atom superposition. This happy coincidence of the largest reduction in error using "single atoms in molecules" is that much of the refinement software in electron and x-ray crystallography that has been developed for free atom atomic scattering factors is largely transferable. This work has been published in two publications in Acta. Cryst. (9,21).

Protein aggregation modeling and experiment

In my newest area of research, we are developing predictive models of protein aggregation. As the competition between aggregation and folding is often decided on time scales shorter than those required for experimental kinetics, experimental insight is difficult to obtain. We have recently extended the protein minimalist model to simulations of multiple chains in obtaining specific structural characteristics of aggregation prone species not observable in the laboratory, focusing on designed sequences for proteins L and G developed in my laboratory, and proposed extensions of these models to new protein sequences of acylphosphatase, a 98 amino acid protein that forms fibrils similar to those found in patients with Alzheimer's disease, and the ordered segment of the prion PrP protein. These models are highly tractable for complete characterization of thermodynamics and kinetics for simulations involving multiple chains to study aggregation propensities, and aggregation abatement by re-engineering protein sequence. These simulations will provide both guidance and feedback from directly related experimental studies of aggregation for these proteins in the Blanch laboratory in Chemical Engineering at UC Berkeley.

Education and Training of Students

Biology and health-related issues have always enjoyed strong public support, while the age of computational sciences and information technology have realized great mass appeal that has transformed human existence in the late half of the last century. The combination of these twin revolutions is inspiring growing numbers of young scientists to enter the rapidly evolving field of computational biology. The popularity of this new area of science has committed faculty like myself to an unusual teaching and education load for an assistant professor, since it is important to spend time introducing, explaining, advising, and advertising this new area of science to students across campus, and not only in bioengineering.

Bioengineering is a new department and much of the curriculum is being developed from the ground up. I have worked with other faculty members in developing a three-to five-year plan for a more cohesive and organized curriculum for bioengineers that will include a common junior year. I have also advised undergraduate students in the BE program every semester, and served on the BE Graduate group admissions committee. As a member of the BE curriculum committee, I am responsible for the development and integration of new computational biology courses in the Bioengineering department, such as

BE131/231: Introduction to Computational Biology (Fall, 2002 and every fall thereafter). Topics include computational approaches and techniques to gene structure and finding, sequence alignment using dynamic programming, protein folding and structure prediction, protein-drug interactions, genetic and biochemical pathways and networks, and microarray analysis. Various "case studies" in these areas are reviewed, web-based computational biology tools will be used by students, and programming projects will be given. Computational biology research connections to biotechnology industry will be explored.

BE143/243: Computational Methods in Biology (Spring 2002, every spring thereafter). - An introduction to biophysical simulation methods and algorithms, including molecular dynamics, Monte Carlo, mathematical optimization, and "non-algorithmic" computation such as neural networks. Various “case studies” in applying these areas in the areas of protein folding, protein structure prediction, drug docking, and enzymatics will be covered.

These courses are attended by not only BE students, but those from computer science, statistics, physics, chemistry and chemical engineering, and molecular and cell biology. I give (standing invitation) guest lectures in CS267:Applications of Parallel Computers and E39B: Introduction to Computational Engineering Science.

Since 2001 I have supervised 4 undergraduates, 4 graduate students, and 7 postdoctoral researchers. (See CV). One graduate student (Dr. Jon Sorenson, chemistry) completed his thesis work with me in 2000. I typically host 3-5 rotation graduate students (in one of three graduate programs) per year.

Local and National Service

At the campus level I have been involved with other faculty in the organization of several major computational biology initiatives. Recently, a large campus-wide effort involving 14 faculty applied for an NIH planning grant for an eventual Program of Excellence in Biomedical Computing (THG lead writer for the education/training component, in addition to the research sections). Computational biology faculty have also organized a response to so-called Tidal Wave II (a strategic plan that addresses the significant increase in student enrollment with increase in faculty retirements at the UC’s over the next decade), and I gladly participated in a formulation of a finalist proposal to the UC Berkeley Strategic planning committee. This year I was head of the BE faculty search committee for Computational Biology.

I am also a member of the COE’s Committee on Computational Engineering Science, and faculty advisor to undergraduate students in this program. I served as Board member for UC Berkeley's SAGE Scholars Program (2001-2002) and was a Regents' and Chancellor's Scholarship Competition Interviewer in 2002. I have been a member of several qualifying exams for students in BE and Chemistry, and conduct preliminary exams in physics for students entering the Biophysics graduate program.

I have served as a panelist for NSF Physical and Theoretical Chemistry Division CAREER awards, NSF Mathematical and Physical Sciences Directorate ADVANCE Fellows, and National Selection Committee for DOE Computational Science Graduate Fellowship Program. I am also a recently appointed Research Council Member of California’s University-Industry Cooperative Research Program. I have organized a number of symposiums and workshops for my scientific community, and have recently been appointed as an Editorial Advisory board member for the Journal of Computational Chemistry, and Editor for Biophysical Journal.

I co-led a nation-wide community effort with John Wooley (associate vice chancellor for research at UC San Diego) in the editing and writing of a white paper on computational biology (see http://cbcg.lbl.gov/ssi-csb and (12)). This paper was circulated among many funding agencies, and provided input into the National Research Council's Frontiers at the Interface of Computing and Biology report in 2001-2002. Most recently I was lead writer for the Biophysical section of DOE/Genomes to Life Goal 4 Roadmap that was presented to a variety of DOE advisory committees, providing information for justifying the FY 2003 budget request and development of the FY 2004 budget submission.

Funding

I currently have support from five research grants: NIH (R01, PPG, and NPEBC), NSF-ITR, and DOE/MICS. I was also given a generous award from the IBM SUR program totaling $750K worth of equipment to purchase a 40-processor IBM/SP machine. I have been the recipient of very generous allocations of cpu (230,000 node hours in FY2002, and 350,000 node hours in FY2003) from the DOE/ERCAP at the NERSC facility. I have a number of pending proposals submitted to ACS-PRF, DOE/BES, and NIH.